COMPARISON OF FEATURE EXTRACTION METHOD OF ROUGH SETS-K-MEANS WITH K-MEANS TO OPTIMIZE VOICE RECOGNITION

Gelar Budidarma Djamal

Program Studi Teknik Informatika
Fakultas Teknik dan Ilmu Komputer Universitas Komputer Indonesia
Jl. Dipati Ukur 114 Bandung
E-mail : gelarbudidarma@gmail.com

ABSTRACT

Speaker verification is the process of voice recognition to verify a speaker. To process speaker verification, voice data must be processed by extraction feature of sound. k-means can be used to increase accuracy in the previous study of voice recognition. The percentage of test results is 79 375%. In addition, there are studies that compare the K-Means, PAM, Rough Sets-K-Means using the datasets of leukemia. The results is the accuracy of Rough Sets-K-Means and PAM higher with k-means. This research will compare rough sets-k-means clustering with k-means to find which suitable algorithm applied to voice recognition.

Before data voice processed, data voice should be extracted using MFCC method. The output of MFCC extraction feature is clustered using rough sets-k-means and k-means. The results feature extraction is classified using LVQ. Testing accuracy using the k-fold cross validation with a value of k = 10.

For test the accuracy using 180 data training and 20 data testing derived from the 5 speakers. The test results show that the accuracy of rough sets-k-means has the highest percentage than k-means in the case of voice recognition, which 80% higher.

Keywords: speaker verification, voice, MFCC, k-means, rough sets-k-means, clustering, learning vector quantization, comparison, k-fold cross validation..

.

1. INTRODUCTION

Speaker verification is the process of voice recognition to verify a speaker. where previously it has been known the identity of the speaker based on the data that has been input. Before data voice processed, data voice should be extracted using MFCC method [1]. Feature extraction that is used in this research is a method of the Mel Frequency Cepstrum Coefficients (MFCC), because the MFCC have advantages in capturing the sound characteristics compared with ZPC (Zero-Crossings with Peak of Amplitudes) and LPC (Linear Predictive Coding) [1].

Previous research that implements the methods of the MFCC in case of Voice Biometrics verification stating the MFCC have a deficiency on the results of the filter of the MFCC because the influence of the length of a sound signal on the parameter values used greatly affect both the poor results from the process of MFCC itself. [2]. Therefore, clustering can make some vector Center as a representative of the whole vector data on the results of parameter MFCC. one that has been tested by using k-means clustering testing only get average percentage rate the accuracy of 79.375% using k-means clustering [3].

In addition there is also a research comparison of K-Means, PAM, Rough Sets-K-Means, with cancer, leukemia, Clustering datasets used for classifying datasets consisting of 3859 genes and other genes, with 50 in because of the complexity of the different leukemia gene pattern, when applied to the k-means algorithm is third, PAM, and rough sets-k-means for classifying leukemia gene datasets, stated that the rough k-means and PAM have better accuracy in gene pattern classifying datasets cancer compared to the k-means [4]. Rough K-Means can be have better accuracy than k-means in the case of optimization of voice recognition.

Based on the description above, this research will be applied in the comparison of Feature extraction methods of Rough Set-K-Means with K-Means voice recognition on a case. Feature extraction stage using MFCC Feature extraction of sound that is as useful to the voice recognition process. Rough Set-K-Means with K-Means is used to optimize results on MFCC filter. Learning Vector Quantization is used as classification in order to get the model pattern. and comparing optimization results the accuracy of K-Means with Rough Sets-K-Means.

2. MODEL ANALYSIS AND DESIGN IMPLEMENTATION

To view a comparison of the use of the k-means method and rough sets-k-means in the case of voice recognition, then the following is the analysis of the process of voice recognition system using the extraction method of the Mel Frequency Cepstrum Coefficients (MFCC), which includes the process of training (training mode) and the process of testing (testing mode).

Training Mode is the part that is in charge of study and model the sound of a speaker. Learning against the speaker's voice is performed by making a model of the speaker's voice then save it to the database.

Testing Mode is the duty to recognize the voice of the speaker. Voice recognition is done by doing a matching test data with the data model of the pattern on the database. In comparison, the matching is done is a 1:1 comparison where the model sounds from A direct comparison with the results of the modeling study of sound that exists in A database. The following process flow voice recognition that can be seen in Figure 1.

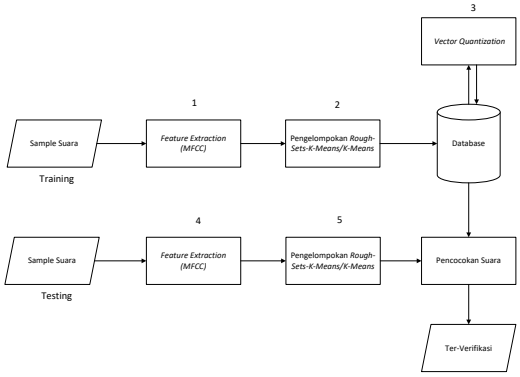


Figure 1. Voice Recognition Process Flow

2.1 Pre Processing

The sound signal will be processed are analog so if will be done digitally processing sound, the signal must be converted into digital signals, in the form of a sequence of numbers with a certain level of precision is called analog to digital conversion using analog-todigital converters (ADC). The concept of working ADC consists of three processes :



Figure 2. The Working Concept Of Analog-Digital Conventer

Description of the working concept of ADC:

1. Sampling a continuous signal conversion is in the domain of time into discrete signals, through a process of sampling a signal at a specified time interval. So that if $x_0(t)$ is the input signal, then the output is $x_0(nT)$, with T is interval sampling.
2. Quantization is the process of data values into numbers gave the number of a certain predetermined in advance.
3. Coding, in this process, each of the discrete values that have been obtained, is represented with n-bit binary number.

2.2 Pre Processing

It is used in the analysis of a sample of the dataset voice consists of 5 speakers, taken within 1 seconds and said that in say is the word "A" in a different intonation.

$$X = FS\ x\ dt(second)\ x\ \left(\frac{bit}{8}\right) x\ j \quad (1)$$

X = The data sampling
FS = The sampling frequency
dt = duration recorded (second)
bit = the number of bits of resolution
j = 1 for *mono*, dan 2 for *stereo*

2.3 MFCC

Method of MFCC consists of 8 process i.e. dc removal, pre-emphasize, frames, windowing, blocking fast fourier transfrom (FFT), the discrete cosine transform, filterbank (DCT) and cepstral liftering process can be seen in Figure 3.

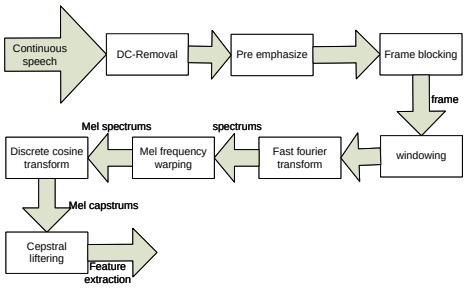


Figure 3. The process method of the Mel Frequency Cepstrum Coefficients (MFCC)

2.3.1 DC Removal

DC Removal Components aim to calculate the average of a data sample sounds, and subtracting the value of each sample with the average value. The goal is get the normalization of data voice input [2].

$$D[n] = s[n] - \bar{x} \ , 0 \leq n \leq N-1 \ (2)$$

D[n] = sample signal DC removal process results
s[n] = original signal sample
$\bar{x}$ = average value of a sample of the original signal.
N = the length of the signal

2.3.2 Pre Emphasize

Pre-emphasize is one type of filter is often used before a signal is processed further. This filter retains the high frequencies in a spectrum, which are eliminated at the time of the sound production process [2].

The most common form used in pre-emphasize filter is as follows:

$$H(z) = 1 - \alpha z^{-1} \quad (3)$$

Where $0.9 \leq \alpha \leq 1.0$, and $\alpha \in r$. Formula above can serve as a first order differentiator, as follows:

$$y[n] = s[n] - \alpha\, s[n-1] \ (4)$$

y[n] = signal after pre-emphasize filter
s[n] = signal before pre-emphasize filter

In general the value of α is most often used is between 0.9 to 1.0.

2.3.3 Frame Blocking

Results of the sound recording is the analog signal in the time domain is time variant, which is a function of the time-dependent. Therefore these signals should be cut into pieces in a particular time slot slot-in order to be considered invariant. The voice signal is cut along the 20 milliseconds. Each piece is called a frame. To count the number of Frames used formulas

The total number of frame = ((I-N)/M)+1) (5)

I = Sample rate
N = Sample point (Sample rate * time framing(s))
M = N/2

2.3.4 Windowing

The process of framing may cause the occurrence of spectral leakage (spectral leakage) or aliasing. Aliasing is a new signal which have different frequencies with the original signal. This effect can occur due to the low number of sampling rate, or because the process frame blocking which causes the signal being discontinued. To reduce the possibility of spectral leakage, then the result of the process of framing must pass a process window [2].

Function on this windowing process using a Hamming Window functions as follows:

$$w(n) = 0.54 - 0.46\cos\frac{2\pi n}{M-1} \ (6)$$

n = 0,1,...,M-1
M = length of the frame

2.3.5 Fast Fourier Transfrom

DFT is an extension of the fourier transform that applies to discrete signals with finite length. All periodic signal of sinusoidal signals combined into one that could be formulated as follows :

$$S[k] = \sum_{n=0}^{N-1} s[n]\, e^{-\frac{j2\pi nk}{N}}, 0 \leq k \leq N-1 \ (7)$$

N = number of samples will be processed ($N \in N$)
S(n) = sample signal value
K = variable-frequency discrete

2.3.6 Filterbank

Filterbank using representation konvolusi in doing a filter against the signal. Konvolusi can be done by doing multiplication between spectrum signal with coefficients filterbank.

The following is the formula used in the calculation of filterbanks. The following calculations to filterbank :

$$Y[i] = \sum_{j=1}^{N} S[j]H_i[j] \ (8)$$

N= total of magnitude spectrum ($N \in N$)
S[j] = magnitude spectrum at frequency j
$H_i[j]$ = filterbank coefficient at frequency j

2.3.7 Discrete Cosine Transfrom

This process is the final step of feature extraction. The result of this is the DCT features required by the author to do the process analysis of the voice recognition. Using the formula:

$$C_n = \sum_{k=0}^{k} (logS_k)\cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{k}\right] \ (9)$$

Dimana
k = coefficient 10
Sk

2.3.8 Cepstral Liftering

The outcome of the main proceedings MFCC feature extraction has some weaknesses. Low order of cepstral coefficients are very sensitive to the spectral slope, while the high order is very sensitive to noise. Therefore, cepstral liftering became one of the standard techniques that are applied to minimize the sensitivity. Cepstral liftering can be done by implementing the functionality of window against cepsral features.

$$w[n] = \left\{1 + \frac{L}{2}\sin\left(\frac{n\pi}{L}\right)\right\} \quad (10)$$

L = number of cepstral coefficients
n = index of cepstral coefficients

2.4 K-Means Clustering

Clustering is the most fundamental factor in pattern recognition. The main problem of clustering is getting some vector value Center that can represent the overall results of the vector feature extraction. K-means clustering is one method used to partition the result vector feature extraction into the k vector Center [5].

K-Means Clustering is the process of mapping vectors that are in scope of the vast territory into a certain number of large (k) vector. The territory is represented by vector quantization process results from the Centre referred to as a cluster. A vector quantization process results from the center known as codewords. While the collection of vector known as Center codebooks [5]

2.5 Rough Sets-K-Means Clustering

Rough Sets-the K-Means algorithm is the use of theory of rough set and k-means algorithm for handling uncertainty not involved in cluster analysis

Step on Rough Sets-K-Means as follows:

1. Select the initial cluster n objects into k clusters.
2. Assign each object on the lower limit (L (x)) or the upper limit (U (x)) on cluster or each cluster as i:

   On each object v, let d (v, xi) to the distance between itself and a centroid cluster on xi, the difference between the d (v, xi)-d (v, xj), $1 \le i, j \le k$ is used to determine a member of v, as follows:

   a. If the d (v, xi)-d (v, xj) ≤ early, then v ∈ U (xi) & v ∈ U (xj). Furthermore, v will not be part of any lower limit.
   b. If not, v ∈ L (x), so d (v, x) is the minimum for $1 \le i \le k$. Additionally, v ∈ U (xi).

$$C_{i=}\begin{cases} W_{lower} \times \frac{\sum_{a\in L(c)} A_j}{|L(c)|} + W_{upper} \times \frac{\sum_{a\in U(c)-L(c)} A_j}{|U(c)-L(c)|} \\ \quad if\ |U(c) - L(c) \neq \phi| \\ W_{lower} \times \frac{\sum_{a\in L(x)} A_j}{|L(C)|} \quad otherwise \end{cases} \quad (11)$$

C(n) = index in centroid
U(c) = centroid Upper
L(c) = centroid Lower
W(lower) = lower bound
W(Upper) = upper bound

3. On each cluster xi recalculate Center according to the following equation the weighted combination of the data point on the

If not where $1 \le j \le k$. Wlower Parameter and with different interests resemble Wupper on lower limit and upper limit
If convergence criteria are met, i.e. the cluster Center similar to the one in the previous iteration, and then stop, go back to step 2.

2.6 Learning Vector Quantization

Learning Vector Quantization (LVQ) is a method for learning or training on a competitive layer. LVQ learning to classify the input vector to the target class that is specified by the user [6].

LVQ network consists of two hidden layers that is competitive and a linear layer layer. Competitive layer also called Self Organizing Folders (SOM). Called competitive because the layer of neurons compete with competition algorithm that will produce winning neuron (winning neuron). On the network, LVQ bias in competitive layer removed. Layers of linear source ragamkan classes on a competitive layer to the target classification specified by the user [6]. LVQ architecture can be seen in Figure 3 here

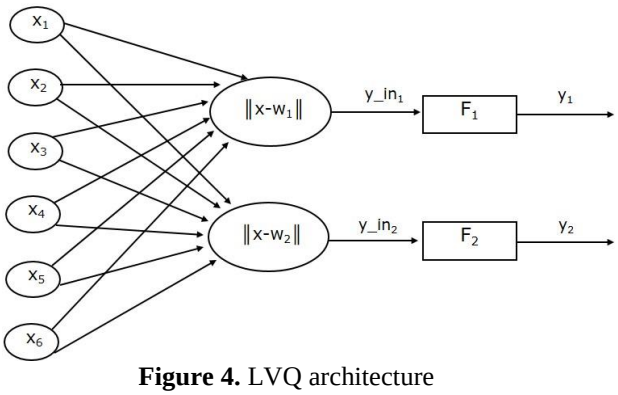


Figure 4. LVQ architecture

2.7 Euclidean Distance

Euclidean distance is a method used to measure the distance (distance). Euclidean distance is actually a generalization of a theorem of phytagoras.

Euclidean distance is formulated in the following equation:

$$d(x, y) = \sqrt{\sum_{i=1}^{r}(x_i - y_i)^2} \quad (12)$$

d = the distance between x and y
r = the number of features in the vector
x_i, y_i = value features to i of x and y

The minimum distance possible between two vector data is 0. In addition, the distance from x to y will equal *y* to *x*, $d(x, y) = d(y, x)$ [5].

3. TESTING

Here is the testing on k-means clustering and rough sets-k-means is done using a k-fold cross validation with a value of k = 10. A value of 10 indicates the number of datasets (D), where a dataset (D) is divided into 10 sections i.e. D1, ... D10 and iteration (testing) done ten times.

On this test used the optimal parameter values which has been done on the test parameter i.e. learning rate α = 0.001, decreased learning rate dec α = 0.001 maximum = 100 epoh and used 180 training data and 20 test data.

3.1 Testing K-Means

Setelah dilakukan pengujian dengan menggunakan *10-fold cross validation* pada *cluster k-means* yang pengujiannya dilakukan sebanyak 10 kali, maka didapatkan rekap hasil pengujian sebagai berikut pada Table 1.

Table 1. The test results on a K-Fold Cross Validation of K-Means

Testing	True	False	Accuracy (%)
1	14	6	70
2	16	4	80
3	15	5	75
4	14	6	70
5	16	4	80
6	14	6	70
7	15	5	75
8	16	4	80
9	15	5	75
10	15	5	75
Average Accuracy (%)			75%

3.1 Testing Rough Sets-K-Means

After the test is performed using a 10-fold cross validation are done performed as many as 10 times, then obtained the following test results recap on Table 2.

Table 2. The test results on a K-Fold Cross Validation Rough Sets-K-Means

Testing	True	False	Accuracy (%)
1	17	3	85
2	17	3	85
3	18	2	90
4	18	2	90
5	15	5	75
6	14	6	70
7	15	5	75
8	17	3	85
9	15	5	75
10	14	6	70
Average Accuracy (%)			80%

4. CONCLUSIONS

4.1 Conclusions

Based on the results of the testing that has been done on this research, then the conclusion that a comparison between the rough sets-k-means and k ¬-means on sound characteristics of MFFC can be inferred by using the value of k (k = 6) and the parameter value to the study of learning vector quantization using the parameters of the learning rate (α = 0.001), decreased learning rate (dec α = 0.1) and maximum epoh = 100 is used because of testing the best accuracy in the use of the datasets that are used but using different parameters can allow can change the result accuracy, so it can be inferred that the rough sets-k-means has better accuracy in the case of voice recognition with a percentage of 80%, and k-means has only a percentage of 75%. Then it can be inferred that the rough sets-k-means more suitable for voice recognition.

4.2 Suggestions

Based on the results of the research that has been done, this research still can be examined further. As for the suggestions for further studies as follows:

1. Extracting the MFCC is used, allowing the extraction of different sound characteristics such as LPC or ZPC would affect a better outcome on the accuracy of voice recognition cases.
2. Noise can affect on the success rate of voice recognition, expected a better method that can address the problem of noise.
3. In the process of learning to be done using other classification methods, such as using Backpropagation Hiden Markov models, Leans

Blue Grey Algorithm which can be used as a comparison in the degree of accuracy voice recognition.

BIBLIOGRAPHY

[1] E. S. Riyanto, Perbandingan Ekstraksi Ciri Suara MFCC, ZCPA dan LPC, STMIK HIMSYA.

[2] Darma Putra Adi Resmawan, "Verifikasi Biometrika Suara Menggunakan Metode MFCC dan DTW," *LONTAR KOMPUTER,* vol. Vol. 2 NO.1, pp. 8-21, 2011.

[3] A. D. Andriana, "Perangkat Lunak Untuk Membuka Aplikasi Pada Komputer Dengan Perintah Suara Menggunakan Metode Mel Frequency Cepstrum Coefficients," *Jurnal Ilmiah Komputer dan Informatika (KOMPUTA),* vol. 2, pp. 21-26, 2013.

[4] Parvesh Kumar, Siri Krishan Wasan, "Comparative Study of K-Means, Pam and Rough K-Means," *2009 International Symposium on Computing, Communication, and Control (ISCCC 2009),* pp. 136-140, 2011.

[5] E. Prasetyo, Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab, Yogyakarta: Andi, 2014.

[6] K. S. M. Ghassemian, Speaker Identification Features, Models and Robustness, Denmark: Kongens Lyngby, 2009.

[7] Mada Sanjaya W.S, Zabidin Salleh, "Implementasi Pengenalan Pola Suara Menggunakan Mel-Frequency Cepstrum Coefficients (MFCC) dan Adaptive Neuro-Fuzzy Inferense System (Anfis) Sebagai Kontrol Lampu Otomatis," *Al-HAZEN Jurnal of Physics,* vol. 1 No. 1, pp. 43-54, 2014.

[8] R. D. Inung Wijayanto, "Jenis Tipe Jangkauan Suara Pada Pria dan Wantia Menggunakan Metoda Mel-Frequency Cepstral Coefficient dan Jaringan Syaraf Tiruan Backpropagation," *Fakultas Elektro dan Komunikasi Insitut Teknologi Telkom Bandung.*

[9] Sugiyono, Metode Penelitian Kuantitatif, Kualitatif, dan Kombinasi (Mixed Method), Bandung: Alfabeta, 2014.

[10] Elsadig Ahmed Mohamed, Hanan Adlan, Abd. Rahman Ramli, "A Combined Rough Sets-K-Means Vector Quantization Model for Arabic Speech Reconition," *(IJCSIT) International Journal of Computer Science and Information Technologies,* vol. 7, pp. 260-265, 2016.